日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年　９月１３日

出願番号
Application Number: 特願２００２－２６９１９３

[ST.10/C]: [ＪＰ２００２－２６９１９３]

出願人
Applicant(s): 富士ゼロックス株式会社

２００３年　５月１６日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０３６２４５

【書類名】　特許願

【整理番号】　FE02-01294

【あて先】　特許庁長官　殿

【提出日】　平成14年 9月13日

【国際特許分類】　G06F 17/00

【発明者】

【住所又は居所】　神奈川県足柄上郡中井町境４３０　グリーンテクなかい
富士ゼロックス株式会社内

【氏名】　劉　紹明

【特許出願人】

【識別番号】　000005496

【氏名又は名称】　富士ゼロックス株式会社

【代理人】

【識別番号】　100098132

【弁理士】

【氏名又は名称】　守山　辰雄

【手数料の表示】

【予納台帳番号】　035873

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9606109

【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　テキスト文比較装置

【特許請求の範囲】

【請求項１】　テキスト文間の比較を行うテキスト文比較装置において、

比較対象となるテキスト文の全体の構造と意味をグラフ理論上の根がある木で表現する木表現手段と、

木表現手段により表現される木の各頂点に単語情報を付与するとともに各辺に単語間の係り受け関係情報である格情報を付与する情報付与手段と、

頂点の間の対応関係及び辺の間の対応関係に基づく木の間の距離を定義する木間距離定義手段と、

木間距離定義手段により定義される木の間の距離を求める木間距離取得手段と、

木の間の距離をテキスト文間の相違を表す距離に適用する木間距離適用手段と、

木間距離適用手段による適用に基づいて、比較対象となるテキスト文間の距離を求めるテキスト文間距離取得手段と、

を備えたことを特徴とするテキスト文比較装置。

【請求項２】　請求項１に記載のテキスト文比較装置において、

木間距離定義手段は、木の各頂点及び各辺にラベルを付けるラベル付け手段と、

木の各頂点及び各辺に番号を付ける番号付け手段と、

頂点間の対応関係及び辺間の対応関係に基づく木の間の写像条件を設定する木間写像条件設定手段と、

木間写像条件設定手段により設定される木の間の写像条件に基づいて木の間の写像を行う木間写像手段と、

頂点間の対応関係及び辺間の対応関係に基づく森の間の写像を行う森間写像手段と、

写像の重みを設定する写像重み設定手段と、

森間の写像と写像の重みに基づく森間の距離を定義する森間距離定義手段と、

出証特２００３－３０３６２４５

木間の写像と写像の重みに基づく木間の距離を定義する距離定義手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項３】　請求項１又は請求項２に記載のテキスト文比較装置において、

木表現手段は、テキスト文全体の構造と意味をグラフ理論上の根があり順序がある木で表現する、

ことを特徴とするテキスト文比較装置。

【請求項４】　請求項１又は請求項２に記載のテキスト文比較装置において、

木表現手段は、テキスト文全体の構造と意味をグラフ理論上の根があり順序がない木で表現する、

ことを特徴とするテキスト文比較装置。

【請求項５】　請求項２に記載のテキスト文比較装置において、

木表現手段は、テキスト文全体の構造と意味をグラフ理論上の根があり順序がある木で表現し、

森間写像手段は、頂点間の対応関係及び辺間の対応関係に基づく順序がある森の間の写像を行い、

森間距離定義手段は、順序がある森の間の写像と写像の重みに基づく順序がある森の間の距離を定義する、

ことを特徴とするテキスト文比較装置。

【請求項６】　請求項２に記載のテキスト文比較装置において、

木表現手段は、テキスト文全体の構造と意味をグラフ理論上の根があり順序がない木で表現する

森間写像手段は、頂点間の対応関係及び辺間の対応関係に基づく順序がない森の間の写像を行い、

森間距離定義手段は、順序がない森の間の写像と写像の重みに基づく順序がない森の間の距離を定義する、

ことを特徴とするテキスト文比較装置。

【請求項７】　請求項１乃至請求項６のいずれか１項に記載のテキスト文比較装置において、

木間距離適用手段は、単語の写像と木の頂点の写像とを対応させる単語写像対応手段と、

格の写像と木の辺の写像とを対応させる格写像対応手段と、

単語の写像の重みを設定する単語写像重み設定手段と、

格の写像の重みを設定する格写像重み設定手段と、

単語写像重みと木の頂点の写像重みとを対応させる単語写像重み対応手段と、

格写像重みと木の辺の写像重みとを対応させる格写像重み対応手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項８】　請求項７に記載のテキスト文比較装置において、

単語写像重み設定手段は、置換と脱落と挿入について、単語の写像の重みを設定し、

格写像重み設定手段は、置換と脱落と挿入について、格の写像の重みを設定し、

単語写像重み対応手段は、単語置換重みと木の頂点間の置換重みとを対応させる単語置換重み対応手段と、単語脱落重みと木の頂点の脱落重みとを対応させる単語脱落重み対応手段と、単語挿入重みと木の頂点の挿入重みとを対応させる単語挿入重み対応手段と、を有し、

格写像重み対応手段は、格置換重みと木の辺の間の置換重みとを対応させる格置換重み対応手段と、格脱落重みと木の辺の脱落重みとを対応させる格脱落重み対応手段と、格挿入重みと木の辺の挿入重みとを対応させる格挿入重み対応手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項９】　請求項７又は請求項８に記載のテキスト文比較装置において、

単語写像重み設定手段は、木の間の写像において２つの頂点が写像した場合に各頂点に格納されている単語の間の置換重みを設定する単語置換重み設定手段と

木の間の写像において頂点が写像できなく脱落された場合に頂点に格納されている単語の脱落重みを設定する単語脱落重み設定手段と、

木の間の写像において頂点が写像できなく挿入された場合に頂点に格納されている単語の挿入重みを設定する単語挿入重み設定手段と、

単語置換重みと単語脱落重みと単語挿入重みとの間の関係を設定する単語写像重み関係設定手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項１０】　請求項７乃至請求項９のいずれか１項に記載のテキスト文比較装置において、

格写像重み設定手段は、木の間の写像において２つの辺が写像した場合に各辺に格納されている格の間の置換重みを設定する格置換重み設定手段と、

木の間の写像において辺が写像できなく脱落された場合に辺に格納されている格の脱落重みを設定する格脱落重み設定手段と、

木の間の写像において辺が写像できなく挿入された場合に辺に格納されている格の挿入重みを設定する格挿入重み設定手段と、

格置換重みと格脱落重みと格挿入重みとの間の関係を設定する格写像重み関係設定手段と、を有する、

ことを特徴とするテキスト文比較装置。

【請求項１１】　請求項９に記載のテキスト文比較装置において、

単語置換重み設定手段は、２つの単語が同一の単語である場合には単語置換重みをゼロと設定し、２つの単語が異なる場合には単語置換重みを正の定数と設定する、

ことを特徴とするテキスト文比較装置。

【請求項１２】　請求項９に記載のテキスト文比較装置において、

単語置換重み設定手段は、単語置換重みを２つの単語の間の距離の値と設定する、

ことを特徴とするテキスト文比較装置。

【請求項１３】　請求項９、請求項１１、又は請求項１２のいずれか１項に

記載のテキスト文比較装置において、

単語脱落重み設定手段は、単語の脱落重みを定数と設定する、

ことを特徴とするテキスト文比較装置。

【請求項１４】　請求項９、請求項１１、又は請求項１２のいずれか１項に記載のテキスト文比較装置において、

単語脱落重み設定手段は、単語の脱落重みを単語の品詞に基づいて設定する、

ことを特徴とするテキスト文比較装置。

【請求項１５】　請求項９又は請求項１１乃至請求項１４のいずれか１項に記載のテキスト文比較装置において、

単語挿入重み設定手段は、単語の挿入重みを定数と設定する、

ことを特徴とするテキスト文比較装置。

【請求項１６】　請求項９又は請求項１１乃至請求項１４のいずれか１項に記載のテキスト文比較装置において、

単語挿入重み設定手段は、単語の挿入重みを単語の品詞に基づいて設定する、

ことを特徴とするテキスト文比較装置。

【請求項１７】　請求項９又は請求項１１乃至請求項１６のいずれか１項に記載のテキスト文比較装置において、

単語写像重み関係設定手段は、「単語脱落重み＋単語挿入重み＞単語置換重み」という関係を設定する、

ことを特徴とするテキスト文比較装置。

【請求項１８】　請求項１０に記載のテキスト文比較装置において、

格置換重み設定手段は、２つの格が同一の格である場合には格置換重みをゼロと設定し、２つの格が異なる場合には格置換重みを正の定数と設定する、

ことを特徴とするテキスト文比較装置。

【請求項１９】　請求項１０に記載のテキスト文比較装置において、

格置換重み設定手段は、全ての格を複数のカテゴリに分類する格分類手段と、格のカテゴリ間の置換重みを設定する格カテゴリ間置換重み設定手段と、を有し、格置換重みを２つの格が属しているカテゴリ間の置換重みと設定する、

ことを特徴とするテキスト文比較装置。

【請求項２０】　請求項１０、請求項１８、又は請求項１９のいずれか１項に記載のテキスト文比較装置において、

格脱落重み設定手段は、格脱落重みを定数と設定する、

ことを特徴とするテキスト文比較装置。

【請求項２１】　請求項１０、請求項１８、又は請求項１９のいずれか１項に記載のテキスト文比較装置において、

格脱落重み設定手段は、格脱落重みを格の種類に基づいて設定する、

ことを特徴とするテキスト文比較装置。

【請求項２２】　請求項１０又は請求項１８乃至請求項２１のいずれか１項に記載のテキスト文比較装置において、

格挿入重み設定手段は、格挿入重みを定数と設定する、

ことを特徴とするテキスト文比較装置。

【請求項２３】　請求項１０又は請求項１８乃至請求項２１のいずれか１項に記載のテキスト文比較装置において、

格挿入重み設定手段は、格挿入重みを格の種類に基づいて設定する、

ことを特徴とするテキスト文比較装置。

【請求項２４】　請求項１０又は請求項１８乃至請求項２３のいずれか１項に記載のテキスト文比較装置において、

格写像重み関係設定手段は、「格脱落重み＋格挿入重み＞格置換重み」という関係を設定する、

ことを特徴とするテキスト文比較装置。

【請求項２５】　請求項１乃至請求項２４のいずれか１項に記載のテキスト文比較装置において、

テキスト文間距離取得手段は、木間距離取得手段により求められる距離の値をテキスト文間の距離とする、

ことを特徴とするテキスト文比較装置。

【請求項２６】　請求項１乃至請求項２４のいずれか１項に記載のテキスト文比較装置において、

テキスト文間距離取得手段は、木間距離取得手段により求められる距離の値を

比較対象となる木の頂点数の和で割り算した結果をテキスト文間の距離とする、

ことを特徴とするテキスト文比較装置。

【請求項２７】　テキスト文間の比較を行うテキスト文比較方法において、

比較対象となるテキスト文の全体の構造と意味をグラフ理論上の根がある木で表現し、

表現される木の各頂点に単語情報を付与するとともに各辺に単語間の係り受け関係情報である格情報を付与し、

頂点の間の対応関係及び辺の間の対応関係に基づく木の間の距離の定義に基づいて、比較対象となるテキスト文の木間の距離を求め、

木の間の距離をテキスト文間の相違を表す距離に適用して、比較対象となるテキスト文間の距離を求める、

ことを特徴とするテキスト文比較方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、例えばコンピュータを利用して、テキスト文間の意味内容の相違を比較する装置及び方法に関し、特に、高精度に実時間で比較を行う装置及び方法に関する。

【０００２】

【従来の技術】

ＩＴ技術、特に高速インターネット・モバイル技術の飛躍的な発展により、大量の情報が誰でも、どこでも、いつでも利用可能になったが、逆に、情報洪水と言われる現象が起き、ユーザが真に必要な情報を取り出すことが困難になってきている。利用者がどのような状況にあっても常に適切な情報を得ることができる世界を実現するために、情報洪水の中からユーザにとって真に価値ある情報を抽出・再構成することが必要である。

【０００３】

ここで、ドキュメントの意味内容の比較や意味内容によるテキスト文書の分類やユーザの情報探索意図の理解に関する技術は重要である。また、ドキュメント

の意味内容の比較や意味内容によるテキスト文書の分類やユーザの情報探索意図の理解を実現するためには、自然言語処理などを利用した意味の類似性判定が欠かせないものである。

【０００４】

この分野では、テキスト文間の類似性を測る技術が幾つか提案されているが、文のローカルな情報、例えば、文に出現している単語情報や、単語間の係り受け情報を利用したものが主流であって、テキスト文の意味内容の評価尺度としては適用しにくく、ドキュメントの意味内容の比較及びユーザの情報探索意図の理解という目標の実現にはつながらなかった。

【０００５】

最近、テキスト文を意味解析してグラフで表現し、グラフ表現の基に経験的類似度を測る手法が提案されたが、提案された類似度には、構造的な変化を考慮していないものもあるし、類似度の定義とテキスト文の意味内容の相違との関係が明確ではないものもあった。

【０００６】

また、本発明と関連する従来技術の例として、以下のようなものがあった。

【０００７】

【非特許文献１】

原田、水野、論文“ＥＤＲを用いた日本語意味解析システムＳＡＧＥ”、「人口知能学会論文誌」、２００１年、１６（１）、ｐ．８５－９３

【非特許文献２】

相澤彰子、論文“語と文書の共起に基づく特徴量の定義と適用”、「自然言語処理」、２０００年３月、１３６－４

【非特許文献３】

馬青、論文“日本語名詞の意味マップの自己組織化”、「情報処理学会論文誌」、２００１年、第４２巻、第１０号

【０００８】

【発明が解決しようとする課題】

上記で述べたように、従来では、テキスト文間の意味内容の類似度を比較する

性能は未だに不十分なものであり、従来提案された類似度はテキスト文間の意味内容の相違の説明につながりにくいといった問題があった。

【０００９】

本発明は、上記のような従来の事情に鑑みなされたもので、テキスト文間の意味内容の相違を高精度に実時間で比較することができる装置や方法を提供することを目的とする。更に具体的には、本発明では、例えば、ドキュメントの意味内容の比較や、意味内容によるテキスト文書の分類や、ユーザの情報探索意図の理解を実現するために、テキスト文間の意味内容の相違を測ることができる距離尺度を数学的に定義して、当該距離尺度を実時間内で求めることを実現する。

【００１０】

【課題を解決するための手段】

上記目的を達成するため、本発明に係るテキスト文比較装置では、次のようにして、テキスト文間の比較を行う。

すなわち、木表現手段が、比較対象となるテキスト文をグラフ理論上の根がある木で表現する。情報付与手段が、木表現手段により表現される木の各頂点に単語情報を付与するとともに各辺に単語間の係り受け関係情報である格情報を付与する。木間距離定義手段が、頂点の間の対応関係及び辺の間の対応関係に基づく木の間の距離を定義する。木間距離取得手段が、木間距離定義手段により定義される木の間の距離を求める。木間距離適用手段が、木間の距離をテキスト文間の相違（或いは、類似）を表す距離に適用する。テキスト文間距離取得手段が、木間距離適用手段による適用に基づいて、比較対象となるテキスト文間の距離を求める。

【００１１】

従って、比較対象となる２つのテキスト文について、テキスト文全体の構造と意味をグラフ理論上の根がある木で表現し、２つの木の間の距離を適用して求められるこれら２つのテキスト文間の距離に基づいて、これら２つのテキスト文間の意味的な相違を検討することができるため、テキスト文間の比較を高精度に実時間で行うことができる。

【００１２】

ここで、本発明では、グラフ理論上の木の間の距離をテキスト文の比較に適用しており、テキスト文に含まれる単語の情報や格の情報ばかりでなく、テキスト文の構造についても考慮している。また、本発明では、木の頂点に単語情報を付与するとともに、木の辺に格情報を付与している。

また、根があり順序がある木を用いるか或いは根があり順序がない木を用いるかによって、テキスト文間の距離を２種類に分けることができ、応用上で計算速度及び比較精度によって任意に選択することができる。

【００１３】

なお、本明細書では、グラフ理論上の根があり順序がある木をＲＯ木（Ｒｏｏｔｅｄ　ａｎｄ　Ｏｒｄｅｒｅｄ　ｔｒｅｅ）と言い、グラフ理論上の根があり順序がない木をＲ木（Ｒｏｏｔｅｄ　ｔｒｅｅ）と言う。

また、ＲＯ木とＲ木とを比較すると、ＲＯ木の方がＲ木と比べて計算が簡易である一方、Ｒ木の方がＲＯ木と比べて一般的に精度がよい。

【００１４】

また、単語情報としては、種々な情報が用いられてもよく、例えば、単語属性の情報が含まれてもよい。単語属性の情報としては、例えば、形態素解析により得られる品詞の情報などを用いることができ、また、動詞の場合には変形の情報などを用いることができる。

また、単語間の係り受けの種類が格に相当する。

また、単語の情報及び格の情報は、例えば、テキスト文を意味解析して求められる。

【００１５】

また、Ｒ木間の写像条件としては、例えば、頂点間の写像と辺間の写像に対して、「１対１写像であり、親子関係（上下関係）を保存し、構造を保存し、頂点間の写像と辺間の写像が交差しない」などの条件が用いられる。

また、ＲＯ木間の写像条件としては、例えば、頂点間の写像と辺間の写像に対して、「１対１写像であり、親子関係（上下関係）を保存し、兄弟間に左右の関係を保存し、構造を保存し、頂点間の写像と辺間の写像が交差しない」などの条件が用いられる。

【００１６】

また、木Ａから木Ｂへ写像するときに、例えば、木Ａの頂点が木Ｂの頂点に写像する場合が頂点の置換に相当し、木Ａにあって写像できない頂点が脱落に相当し、木Ｂにあって写像できない頂点が挿入に相当する。

また、木と木の間の距離としては、例えば、１本の木を他の１本の木へ変換するときの重みの和（写像重みの和）の最小値が用いられる。また、このような木と木の間の距離には、暗黙的に、森と森の間の距離が含まれる。

【００１７】

また、ＲＯ木或いはＲ木の各頂点に番号を付ける仕方としては、例えば、深さ優先探索により各頂点に番号を増大順で付けて、番号が大きい頂点から順に計算を行う仕方が用いられる。具体的には、動的計画法を用いて、一番下側の部分木から上側の部分木へと順に計算を行っていく。

また、ラベルは、情報を記憶するためのものである。

【００１８】

以下で、更に、本発明の構成例を示す。

（１）テキスト文間の意味内容を測る距離を求めるテキスト文間の意味内容比較装置において、テキスト文全体の構造と意味をグラフ理論上のＲＯ木或いはＲ木で表現する手段と、ＲＯ木或いはＲ木の各頂点と各辺にそれぞれ単語情報と関連する単語間の係り受け関係（格）情報を付与する手段と、頂点間及び辺間の対応関係に基づくＲＯ木間或いはＲ木間の距離を定義する手段と、定義したＲＯ木間或いはＲ木間の距離を求める手段と、ＲＯ木間或いはＲ木間の距離をテキスト文間の意味的な相違を比較する距離に適用する手段と、テキスト文間の距離を求める手段を備える。

【００１９】

（２）頂点間及び辺間の対応関係に基づくＲＯ木間或いはＲ木間の距離を定義する手段は、ＲＯ木或いはＲ木の各頂点と各辺にそれぞれラベルを付けるラベル付け手段と、ＲＯ木或いはＲ木の各頂点と各辺にそれぞれ番号を付ける番号付け手段と、頂点間と辺間の対応関係に基づくＲＯ木間の写像条件と、頂点間と辺間の対応関係に基づくＲ木間の写像条件と、頂点間と辺間の対応関係に基づくＲＯ

木間の写像手段と、頂点間と辺間の対応関係に基づくＲ木間の写像手段と、頂点間と辺間の対応関係に基づく順序がある森の間の写像を行う写像手段と、頂点間と辺間の対応関係に基づく順序がない森の間の写像を行う写像手段と、これらの写像手段による写像の重みを定義する写像重み設定手段と、順序がある森の間の写像手段と写像重み設定手段に基づく順序がある森の間の距離を定義する手段と、順序がない森の間の写像手段と写像重み設定手段に基づく順序がない森の間の距離を定義する手段と、ＲＯ木間の写像手段と写像重み設定手段に基づくＲＯ木間の距離を定義する手段と、Ｒ木間の写像手段と写像重み設定手段に基づくＲ木間の距離を定義する手段を有する。

【００２０】

（３）ＲＯ木間或いはＲ木間の距離をテキスト文間の意味的な相違を測る距離に適用する手段は、単語間の写像をＲＯ木或いはＲ木の頂点間の写像に対応させる手段と、格間の写像をＲＯ木或いはＲ木の辺間の写像に対応させる手段と、単語間の写像重みをＲＯ木或いはＲ木の頂点間の写像重みに対応させる手段と、格間の写像重みをＲＯ木或いはＲ木の辺間の写像重みに対応させる手段と、単語の写像重みを設定する手段と、格の写像重みを設定する手段を有する。

【００２１】

（４）テキスト文間の距離を求める手段は、ＲＯ木間或いはＲ木間の距離を求める手段で求められた距離値をテキスト文間の距離とする。

（５）テキスト文間の距離を求める手段は、ＲＯ木間或いはＲ木間の距離を求める手段で求められた距離値と２つのＲＯ木或いはＲ木の頂点数の和とを割り算した結果をテキスト文間の距離尺度とする。

【００２２】

（６）単語間の写像重みを設定する手段は、ＲＯ木間或いはＲ木間の写像において、２つの頂点が写像した場合には、各頂点に格納されている単語間の置換重みを設定する手段と、頂点が写像できなく、脱落された場合には、頂点に格納されている単語の脱落重みを設定する手段と、頂点が写像できなく、挿入された場合には、頂点に格納されている単語の挿入重みを設定する手段と、単語置換重みと単語脱落重みと単語挿入重みとの間の関係を設ける手段を有する。

【００２３】

（７）格の写像重みを設定する手段は、ＲＯ木間或いはＲ木間の写像において、２つの辺が写像した場合には、各辺に格納されている格間の格置換重みを設定する手段と、辺が写像できなく、脱落された場合には、辺に格納されている格の脱落重みを設定する手段と、辺が写像できなく、挿入された場合には、辺に格納されている格の挿入重みを設定する手段と、格置換重みと格脱落重みと格挿入重みとの間の関係を設ける手段を有する。

【００２４】

（８）単語置換重みを設定する手段は、２つの単語が同一の単語である場合には単語の置換重みをゼロと設定する手段、２つの単語が異なる場合には単語の置換重みを正の定数と設定する手段を有する。

（９）単語置換重みを設定する手段は、単語の置換重みを単語間の距離と設定する。

【００２５】

（１０）単語脱落重みを設定する手段は、単語の脱落重みを定数と設定する。

（１１）単語脱落重みを設定する手段は、単語の品詞によって単語の脱落重みを求める。

（１２）単語挿入重みを設定する手段は、単語の挿入重みを定数と設定する。

（１３）単語挿入重みを設定する手段は、単語の品詞によって単語の挿入重みを求める。

（１４）単語の置換重み、脱落重み、挿入重みの間の関係を設ける手段は、「単語脱落重み＋単語挿入重み＞単語置換重み」を満たす関係を設ける。

【００２６】

（１５）格置換重みを設定する手段は、２つの格が同一の格である場合には格の置換重みをゼロと設定する手段と、２つの格が異なる場合には格の置換重みを正の定数と設定する手段を有する。

（１６）格置換重みを設定する手段は、全ての格を複数であるＮ個のカテゴリに分類する手段と、格カテゴリ間の置換重みを設定する手段と、格間の置換重みを、２つの格が属しているカテゴリ間の置換重みとする手段を有する。

【００２７】

（１７）格脱落重みを設定する手段は、格の脱落重みを定数と設定する。

（１８）格脱落重みを設定する手段は、格の種類によって格の脱落重みを求める。

（１９）格挿入重みを設定する手段は、格の挿入重みを定数と設定する。

（２０）格挿入重みを設定する手段は、格の種類によって格の挿入重みを求める。

（２１）格の置換重み、脱落重み、挿入重みの間の関係を設ける手段は、「格脱落重み＋格挿入重み＞格置換重み」を満たす関係を設ける。

【００２８】

（２２）テキスト文間の意味内容を測る距離を求めるテキスト文間の意味内容比較方法において、テキスト文全体の構造と意味をグラフ理論上のＲＯ木或いはＲ木で表現し、ＲＯ木或いはＲ木の各頂点と各辺にそれぞれ単語情報と単語間の係り受け関係（格）情報を格納し、頂点間及び辺間の対応関係に基づくＲＯ木間或いはＲ木間の距離を定義した内容に基づいて、定義したＲＯ木間或いはＲ木間の距離を求め、ＲＯ木間或いはＲ木間の距離をテキスト文間の意味的な相違を比較する距離に適用して、テキスト文間の距離を求める。

【００２９】

【発明の実施の形態】

本発明に係る一実施例を図面を参照して説明する。

図１には、本発明の一実施例に係るテキスト文間の意味内容の比較装置（テキスト文比較装置）の実施の一形態を示してあり、当該装置は本発明の一実施例に係るテキスト文の意味内容の比較方法を実施する。

【００３０】

同図に示したテキスト文比較装置には、外部記憶装置１と、テキスト文の形態素を抽出する形態素解析部２と、テキスト文の意味を解析する意味解析部３と、意味解析部３で解析した結果をグラフ理論上のＲＯ木或いはＲ木に変換する木構造変換部４と、２つの単語を置換するときの単語置換重みと単語を脱落するときの単語脱落重みと単語を挿入するときの単語挿入重みを求める単語写像重み計算

部５と、２つの格を置換するときの格置換重みと格を脱落するときの格脱落重みと格を挿入するときの格挿入重みを求める格写像重み計算部６と、ＲＯ木間或いはＲ木間の距離を計算する距離計算部７と、テキスト文間の意味内容の相違を求める意味内容比較部８と、例えばメモリから構成される記憶部９と、複数のメモリ１０～１８が備えられている。

【００３１】

なお、外部記録装置１には、テキスト文のデータが格納されている。

また、メモリ１０とメモリ１１は、外部記憶装置１から読み出した２つのテキスト文のデータをそれぞれ記憶する。メモリ１２とメモリ１３は、２つのテキスト文のそれぞれについて、形態素解析部２による解析結果を記憶する。メモリ１４とメモリ１５は、２つのテキスト文のそれぞれについて、意味解析部３による意味解析の結果を記憶する。メモリ１６とメモリ１７は、２つのテキスト文のそれぞれについて、木構造変換部４による変換結果を記憶する。メモリ１８は、距離計算部８で求められたＲＯ木間或いはＲ木間の距離を記憶する。

なお、これらのメモリ１０～１８を１つにまとめた構成や、或いは、これらのメモリ１０～１８を設けないような構成が用いられてもよい。

【００３２】

形態素解析部２は、メモリ１０とメモリ１１に格納されている２つのテキスト文の形態素及び属性を抽出し、それぞれのテキスト文の解析結果をそれぞれのメモリ１２、メモリ１３に記憶させる。

意味解析部３は、メモリ１２とメモリ１３に記録されている形態素解析の結果を入力し、テキスト文の意味を解析することを行い、それぞれのテキスト文の解析結果をそれぞれのメモリ１４、メモリ１５に格納する。

【００３３】

木構造変換部４は、メモリ１４とメモリ１５に格納されている意味解析の結果をＲＯ木或いはＲ木に変換し、変換したＲＯ木或いはＲ木の頂点にテキスト文に出現した単語（単語の属性を含む）情報を記憶させるとともに辺にテキスト文に出現した関連する格情報を記憶させる。

また、木構造変換部４は、それぞれのテキスト文について変換した結果をそれ

ぞれのメモリ１６、メモリ１７に格納する。

【００３４】

単語写像重み計算部５は、ＲＯ木間或いはＲ木間の距離を求めるときに必要となる単語置換重み、単語脱落重み、単語挿入重みを求めて、距離計算部７に提供する。

格写像重み計算部６は、ＲＯ木間或いはＲ木間の距離を求めるときに必要となる格置換重み、格脱落重み、格挿入重みを求めて、距離計算部７に提供する。

【００３５】

距離計算部７は、メモリ１６とメモリ１７に格納されている２つのＲＯ木間或いはＲ木間の距離を求め、その結果をメモリ１８に記憶させる。

意味内容比較部８は、メモリ１８に記憶されているＲＯ木間或いはＲ木間の距離を利用して、テキスト文間の距離を求め、その結果を記憶部９に格納する。

【００３６】

次に、本発明に係るテキスト文間の意味内容を比較する距離を計算する装置及び方法の適用例として、情報端末装置に適用した場合の装置構成例を示す。

図２には、本発明に係るテキスト文間の意味内容を比較する距離を計算する方法を情報端末装置に適用した場合の装置構成例を示してある。

同図に示した情報端末装置２０は、外部記憶装置２１と、キーボート２２と、ディスプレイ２３と、プロセッサ部２４から構成されている。プロセッサ部２４には、テキスト文間の距離を求めるモジュール２５が備えられている。

【００３７】

外部記憶装置２１は、入力されたテキスト文のデータや、単語写像重みを求めるために用いられる単語特徴量辞書或いはシソーラス辞書や、格写像重みを求めるために用いられる重み辞書などや、求められたテキスト文間の距離の結果や、ソフトウエアなどを格納し、また、計算に用いられる記憶空間として機能する。ここで、単語特徴量辞書やシソーラス辞書や重み辞書などは、例えば、予め作成され或いは既存のものが用意される。また、具体的に、外部記憶装置２１としては、例えばハードディスクなどで構成することができる。

【００３８】

キーボート２２は、ユーザが操作を指示するための入力装置である。なお、他の入力装置が付加されていてもよい。

ディスプレイ２３は、ユーザに対するメッセージやテキスト文のデータや、解析結果や、距離の計算結果などを表示するための出力装置である。なお、他の出力装置が付加されてもよい。

【００３９】

プロセッサ部２４は、外部記憶装置２１に格納されているソフトウエアなどに従って、実際の処理を行う。具体的に、プロセッサ部２４としては、例えば、マイクロプロセッサや、パーソナルコンピュータなどのコンピュータシステムで構成することができる。そして、上記図１に示した形態素解析部２や、意味解析部３や、木構造変換部４や、単語写像重み計算部５や、格写像重み計算部６や、距離計算部７や、意味内容比較部８は、このプロセッサ部２４の上で動作するソフトウエアによって構成することができる。

【００４０】

次に、本発明の一実施例に係るテキスト文間の意味内容の相違を比較する装置の動作を更に詳細に説明する。

外部記憶装置１には、テキスト文のデータを格納している。外部記憶装置１から２つのテキスト文のデータを読み出し、メモリ１０とメモリ１１にそれぞれ記憶させる。形態素解析部２は、メモリ１０とメモリ１１に記憶しているテキスト文の形態素を抽出し、その結果をそれぞれメモリ１２とメモリ１３に格納させる。

【００４１】

ここで、形態素解析ツールとしては、公表された任意のものを利用することができ、例えば、奈良先端技術大学院大学の松本研究室により公表された“茶筅”形態素解析ツールを用いることができる。

また、図３には、テキスト文「先生は生徒に英語を教える」についての形態素解析の解析結果を示してある。

【００４２】

意味解析部３は、メモリ１２とメモリ１３に記憶された形態素解析の結果を入

力し、テキスト文の構文や、係り受け関係（格）や、テキスト文の深層構造などを解析し、解析した結果をそれぞれメモリ１４とメモリ１５に格納する。

ここで、意味解析ツールとしては、公表された任意の意味解析ツールを利用することができ、例えば、非特許文献１に記載された方法を用いることができる（非特許文献１参照。）。

【００４３】

木構造変換部４は、メモリ１４とメモリ１５に記憶された解析結果を入力し、テキスト文を木構造へ変換して、変換した木構造をそれぞれメモリ１６とメモリ１７に格納する。

図４には、テキスト文「先生は生徒に英語を教える」についての意味解析の解析結果を木構造の形に書き換えたものを示してある。単語情報として、「先生」、「英語」、「生徒」、「に」、「教える」が各頂点に格納されており、格情報として、「先生」と「教える」の間の「ＳＵＢＪ」、「英語」と「教える」の間の「ＯＢＪ」、「生徒」と「に」の間の「ＯＢＪ」、「に」と「教える」の間の「ＯＢＬ」が各辺に格納されている。

【００４４】

上記図４において、格情報として、ＳＵＢＪ（主格）、ＯＢＪ（目的格）、ＯＢＬ（任意格）を示してある。また、格情報として、ＡＤＪＵＮＣＴ（付加格）などを用いることもできる。

なお、本例では、ＯＢＬについては、格助詞と、言い換え可能な格助詞の数だけ変数を用意している。例えば、「彼は京都［に／へ］行った。」の場合、「に」と「へ」が言い換え可能なので、この変数名をＯＢＬ＿ｎｉ－ｈｅとする。

【００４５】

単語写像重み計算部５は、単語置換重み、単語脱落重み、単語挿入重みを求めて、距離計算部７に提供する。

単語置換重みとしては、定数と設定する態様や、単語間の距離を用いて設定する態様を用いることができる。前者の態様では、２つの単語が同じ単語である場合には単語置換重みをゼロと設定し、そうではない場合には単語置換重みを正の定数と設定する。後者の態様では、単語置換重み計算部５は、２つの単語間の距

離を求め、その距離の値を単語置換重みとして、距離計算部７に提供する。

【００４６】

ここで、単語間の距離を求める方法としては、公開された任意の方法を利用することができ、例えば、統計的な方法や、シソーラス辞書を用いた方法や、ニューラルネットを用いた方法がある。統計的な手法としては、例えば、非特許文献２に記載されたＴＦ／ＩＤＦ方法により求めることができる（非特許文献２参照。）。シソーラス辞書を用いた手法としては、例えば、２つの単語が属している概念間の最短道の長さを単語間の距離とすることができる。ニューラルネットワークを用いた手法としては、例えば、非特許文献３に記載された方法を利用することができる（非特許文献３参照。）。また、他の公開された方法を利用することもできる。

【００４７】

単語脱落重みとしては、定数と設定する態様や、単語の品詞情報によって単語脱落重みを設定する態様を用いることができる。後者の態様では、単語の品詞に重みを付け、単語脱落重みを品詞重みと定数との積と設定する。品詞重みの設定としては、例えば、重要な役割を有する品詞に大きな重みを付与する仕方を用いることが好ましく、一例として、動詞の重みが一番重たく、形容動詞、名詞、副詞、形容詞などの順で品詞重みを軽くするように設定することができる。また、他の順番で品詞重みを設定することもできる。

【００４８】

単語挿入重みとしては、定数と設定する態様や、単語の品詞情報によって単語挿入重みを設定する態様を用いることができる。後者の態様では、単語の品詞に重みを付け、単語挿入重みを品詞重みと定数との積と設定する。品詞重みの設定としては、上記した単語脱落重みに関して述べた品詞重みの設定方法と同様な方法で設定することができ、また、異なる方法で設定することもできる。

【００４９】

格置換重み計算部６は、格置換重み、格脱落重み、格挿入重みを求めて、距離計算部７に提供する。

格置換重みとしては、定数と設定する態様や、格の間の距離を用いて設定する

態様を用いることができる。前者の態様では、２つの格が同じ格である場合には格置換重みをゼロと設定し、そうではない場合には各置換重みを正の定数と設定する。後者の態様では、格置換重み計算部６は、２つの格間の距離を求め、その距離の値を格置換重みとして、距離計算部７に提供する。

【００５０】

ここで、格間の距離を求める方法の一例を示す。

まず、全ての格をその内容によって幾つかのカテゴリに分類する。なお、カテゴリの要素数は１以上である。

また、図５に示されるような格カテゴリ間の距離の表を用意しておく。同図に示される表では、複数であるｍ個の格カテゴリの全ての組み合わせについて、格カテゴリ間の距離（距離１１～距離ｍｍ）が設定されている。

次に、与えられた２つの格情報により特定される２つの格が属している格カテゴリをそれぞれ求め、上記図５に示される格カテゴリ間の距離表を用いて当該求められた２つの格カテゴリ間の距離値を求め、当該求められた距離値を２つの格間の距離とする。

なお、格間の距離を求める方法として、他の方法が用いられてもよい。

【００５１】

格脱落重みとしては、定数と設定する態様や、格の種類によって格脱落重みを設定する態様を用いることができる。後者の態様では、格に重みを付け、格脱落重みを格重みと定数との積と設定する。格重みの設定としては、例えば、ＳＵＢＪの重みが一番重たく、ＯＢＪ、ＯＢＬ、ＡＤＪＵＮＣＴなどの順で重みを軽くするように設定することができる。また、他の順番で格重みを設定することもできる。

【００５２】

格挿入重みとしては、定数と設定する態様や、格の種類によって格挿入重みを設定する態様を用いることができる。後者の態様では、格に重みを付け、格挿入重みを格重みと定数との積と設定する。格重みの設定としては、上記した格脱落重みに関して述べた格重みの設定方法と同様な方法で設定することができ、また、異なる方法で設定することもできる。

【００５３】

距離計算部７は、メモリ１６とメモリ１７に記憶されたＲＯ木間或いはＲ木間の距離を求め、その結果をメモリ１８に格納する。

ここで、頂点と辺の対応関係に基づくＲＯ木間の距離は、例えば、特願２００２－０７１２７３号に記載された方法で求めることができる。また、頂点と辺の対応関係に基づくＲ木間の距離は、例えば、特願２００２－０７１２７４号に記載された方法で求めることができる。

【００５４】

次に、上記した特願２００２－０７１２７３号に記載されたＲＯ木間の距離を求める方法や、上記した特願２００２－０７１２７４号に記載されたＲ木間の距離を求める方法を示す。

まず、木の間の距離を記述するために、関連する記号を定義する。

ＲＯ木或いはＲ木T_aの頂点ｘを根とする部分木を$T_a(x)$で表す。

部分木$T_a(x)$の頂点の集合を$V_a(x)$で表す。

部分木$T_a(x)$の辺の集合を$E_a(x)$で表す。

頂点ｘの子供をx_1、x_2、・・・、x_mとし、頂点ｘの子供の集合をＣｈ（ｘ）で表す。

また、本明細書では、次のように定義する。

【００５５】

【数１】

部分木$T_a(x)$と辺$\tilde{x}$からなる部分を$\tilde{T}_a(x)$で表し、$\tilde{T}_a(x)$も部分木と呼ぶ。

部分木$\tilde{T}_a(x)$の頂点の集合を$\tilde{V}_a(x)$で表し、部分木$\tilde{T}_a(x)$の辺の集合を$\tilde{E}_a(x)$で表す。この場合、$V_a(x)=\tilde{V}_a(x)$となり、$E_a(x)\cup\{\tilde{x}\}=\tilde{V}_a(x)$となる。

部分木$\tilde{T}_a(x_1)$、$\tilde{T}_a(x_2)$、・・・、$\tilde{T}_a(x_m)$からなる部分を森と呼び、森を$\tilde{F}_a(x)$で表す。

【００５６】

また、上述したテキスト文の木構造表現法と変換方法から分かるように、頂点ｘ、ｙはテキスト文に出現している単語（単語の属性を含む）を表している。また、関数δ（ｘ，ｙ）は頂点の置換重みを表すとし、単語の置換重みで求めることができる。また、ｑ（ｘ）は頂点ｘの挿入重みを表すとし、単語の挿入重みで求めることができる。また、ｒ（ｘ）は頂点ｘの脱落重みを表すとし、単語の脱落重みで求めることができる。

また、辺に関して、次のように定義する。

【００５７】

【数２】

上述したテキスト文の木構造表現法と変換方法から分かるように、辺 $\tilde{x}$ 、$\tilde{y}$ はテキスト文の単語間の係り受け関係（格）の情報を表している。また、関数 $\delta(\tilde{x},\tilde{y})$ は辺の置換重みを表すとし、格の置換重みで求めることができる。また、$q(\tilde{x})$ は辺 $\tilde{x}$ の挿入重みを表すとし、格の挿入重みで求めることができる。また、$r(\tilde{x})$ は辺 $\tilde{x}$ の脱落重みを表すとし、格の脱落重みで求めることができる。

【００５８】

初めに、上記した特願２００２－０７１２７３号に記載された頂点と辺の対応関係に基づくＲＯ木間の距離を求める方法を示す。

まず、ＲＯ木の根から深さ優先順で頂点と辺に番号を付ける。大きい番号を根とするＲＯ木の部分から小さい番号を根とするＲＯ木の部分の順で部分木間の距離を求めて、最後に全体的なＲＯ木間の距離を求める。

【００５９】

図６（ａ）～図６（ｄ）には、例えばＲＯ木である２つの部分木を４つの態様について示してある。

ＲＯ木である部分木間の距離及び順序がある森間の距離に関して、次のように定義する。

【００６０】

【数３】

図６（ａ）に示されるＲＯ木である２つの部分木$T_a(x)$、$T_b(y)$の間の距離を$D(T_a(x),T_b(y))$で表す。

図６（ｂ）に示されるＲＯ木である２つの部分木$\widetilde{T}_a(x)$、$\widetilde{T}_b(y)$の間の距離を$D(\widetilde{T}_a(x),\widetilde{T}_b(y))$で表す。

図６（ｃ）に示されるＲＯ木である２つの部分木$\widetilde{T}_a(x)$、$T_b(y)$の間の距離を$D(\widetilde{T}_a(x),T_b(y))$で表す。

図６（ｄ）に示されるＲＯ木である２つの部分木$T_a(x)$、$\widetilde{T}_b(y)$の間の距離を$D(T_a(x),\widetilde{T}_b(y))$で表す。

また、順序がある森間の距離$D(\widetilde{F}_a(x),\widetilde{F}_b(y))$、全ての部分木間の距離$D(T_a(x_i),T_b(y))$、$D(\widetilde{T}_a(x_i),\widetilde{T}_b(y))$、$D(\widetilde{T}_a(x_i),T_b(y))$、$D(T_a(x_i),\widetilde{T}_b(y))$、$D(T_a(x),T_b(y_j))$、$D(\widetilde{T}_a(x),\widetilde{T}_b(y_j))$、$D(\widetilde{T}_a(x),T_b(y_j))$、$D(T_a(x),\widetilde{T}_b(y_j))$は既に求められたものとする。

【００６１】

式１～式４を用いて、同図（ａ）～同図（ｄ）のそれぞれに示される２つのＲＯ木の間の距離を求めることができる。ここで、式１～式４中で記号“Ａ－Ｂ”は集合Ａから集合Ｂの全ての要素を取り除く関数を表す。

【００６２】

【数４】

$$D(T_a(x),T_b(y))=\min\begin{cases}\delta(x,y)+D(\widetilde{F}_a(x),\widetilde{F}_b(y)),\\ \min_{y\in CH(x)}\left\{D(\widetilde{T}_a(x),T_b(y))+\sum r(k)\mid k\in(V_a(x)-V_a(x))+\sum r(\tilde{k})\mid \tilde{k}\in(E_a(x)-\widetilde{E}_a(x))\right\},\\ \min_{y\in CH(y)}\left\{D(T_a(x),\widetilde{T}_b(y))+\sum q(k)\mid k\in(V_b(y)-V_b(y))+\sum q(\tilde{k})\mid \tilde{k}\in(E_b(y)-\widetilde{E}_b(y))\right\},\end{cases}$$

・・（式１）

【００６３】

【数５】

$$D(\tilde{T}_a(x),\tilde{T}_b(y))=\min\begin{cases}\delta(x,y)+\delta(\tilde{x},\tilde{y})+D(\tilde{F}_a(x),\tilde{F}_b(y)),\\ \min_{x_i\in Ch(x)}\left\{\begin{matrix}\min\{D(T_a(x_i),T_b(y))+\delta(\tilde{x},\tilde{y}),\quad D(\tilde{T}_a(x_i),\tilde{T}_b(y))\}\\ +\sum r(k)\,|\,k\in(\tilde{V}_a(x)-\tilde{V}_a(x_i))+\sum r(\tilde{k})\,|\,\tilde{k}\in(\tilde{E}_a(x)-\tilde{E}_a(x_i))\end{matrix}\right\},\\ \min_{y_j\in Ch(y)}\left\{\begin{matrix}\min\{D(T_a(x),T_b(y_j))+\delta(\tilde{x},\tilde{y}),\quad D(\tilde{T}_a(x),\tilde{T}_b(y_j))\}\\ +\sum q(k)\,|\,k\in(\tilde{V}_b(y)-\tilde{V}_b(y_j))+\sum q(\tilde{k})\,|\,\tilde{k}\in(\tilde{E}_b(y)-\tilde{E}_b(y_j))\end{matrix}\right\},\end{cases}$$

・・（式２）

【００６４】

【数６】

$$D(\tilde{T}_a(x),T_b(y))=\min\begin{cases}\delta(x,y)+\tilde{r}+D(\tilde{F}_a(x),\tilde{F}_b(y)),\\ \min_{x_i\in Ch(x)}\{D(\tilde{T}_a(x_i),T_b(y))+\sum r(k)\,|\,k\in(V_a(x)-V_a(x_i))+\sum r(\tilde{k})\,|\,\tilde{k}\in(\tilde{E}_a(x)-\tilde{E}_a(x_i))\},\\ \min_{y_j\in Ch(y)}\{D(\tilde{T}_a(x),\tilde{T}_b(y_j))+\sum q(k)\,|\,k\in(V_b(y)-V_b(y_j))+\sum q(\tilde{k})\,|\,\tilde{k}\in(E_b(y)-\tilde{E}_b(y_j))\}\end{cases}$$

・・（式３）

【００６５】

【数７】

$$D(T_a(x),\tilde{T}_b(y))=\min\begin{cases}\delta(x,y)+\tilde{q}+D(\tilde{F}_a(x),\tilde{F}_b(y)),\\ \min_{x_i\in Ch(x)}\{D(\tilde{T}_a(x_i),\tilde{T}_b(y))+\sum r(k)\,|\,k\in(V_a(x)-V_a(x_i))+\sum r(\tilde{k})\,|\,\tilde{k}\in(E_a(x)-\tilde{E}_a(x_i))\},\\ \min_{y_j\in Ch(y)}\{D(T_a(x),\tilde{T}_b(y_j))+\sum q(k)\,|\,k\in(V_b(y)-V_b(y_j))+\sum q(\tilde{k})\,|\,\tilde{k}\in(E_b(y)-\tilde{E}_b(y_j))\}\end{cases}$$

・・（式４）

【００６６】

図７には、例えば順序がある２つの森を示してある。式５を用いて、これら２つの森の間の距離を求めることができる。ここで、記号｜Ａ｜は集合Ａの要素数

を表す。

【００６７】

【数８】

同図に示される２つの順序がある森$\tilde{F}_a(x)$、$\tilde{F}_b(x)$の間の距離$D(\tilde{F}_a(x),\tilde{F}_b(y))$が次のように求められる。

(5-1) 境界条件($1 \le i \le |Ch(x)|$, $1 \le j \le |Ch(y)|$)

$$d1(0,0)=0;$$

$$d1(i,0)=d1(i-1,0)+\sum r(k)\,|\,k\in V_a(x)+\sum r(\tilde{k})\,|\,\tilde{k}\in\tilde{E}_a(x);$$

$$d1(0,j)=d1(0,j-1)+\sum q(k)\,|\,k\in V_b(y)+\sum q(\tilde{k})\,|\,\tilde{k}\in\tilde{E}_b(y);$$

(5-2) $d1(i,j)$の計算($1 \le i \le |Ch(x)|$, $1 \le j \le |Ch(y)|$)

$$d1(i,j)=\min\begin{Bmatrix} d1(i-1,j-1)+D(\tilde{T}_a(x),\tilde{T}_b(y)), \\ d1(i,j-1)+\sum q(k)\,|\,k\in V_b(y)+\sum q(\tilde{k})\,|\,\tilde{k}\in\tilde{E}_b(y), \\ d1(i-1,j)+\sum r(k)\,|\,k\in V_a(x)+\sum r(\tilde{k})\,|\,\tilde{k}\in\tilde{E}_a(x) \end{Bmatrix};$$

(5-3) $$D(\tilde{F}_a(x),\tilde{F}_b(y))=d1(|Ch(x)|,|Ch(y)|).$$

・・（式５）

【００６８】

なお、上記式１について、頂点ｘが葉（Ｃｈ（ｘ）＝ＮＵＬＬ：空集合）である場合には、明らかに、上記式１の右側の第２項を計算する必要がないため、式６を用いて距離Ｄ（$Ｔ_a$（ｘ），$Ｔ_b$（ｙ））を求めることができる。

また、上記式１について、頂点ｙが葉（Ｃｈ（ｙ）＝ＮＵＬＬ：空集合）である場合には、明らかに、上記式１の右側の第３項を計算する必要がないため、式７を用いて距離Ｄ（$Ｔ_a$（ｘ），$Ｔ_b$（ｙ））を求めることができる。

【００６９】

【数９】

$$D(T_1(x),T_2(y))=\min\begin{cases}\delta(x,y)+D(\tilde{F}_1(x),\tilde{F}_2(y)),\\ \min_{y\in Ch(y)}\{D(T_1(x),\tilde{T}_2(y))+\sum q(k)\mid k\in(V_2(y)-V_2(y))+\sum q(\tilde{k})\mid \tilde{k}\in(E_2(y)-\tilde{E}_2(y))\}\end{cases}$$

・・（式６）

【００７０】

【数１０】

$$D(T_1(x),T_2(y))=\min\begin{cases}\delta(x,y)+D(\tilde{F}_1(x),\tilde{F}_2(y)),\\ \min_{x\in Ch(x)}\{D(\tilde{T}_1(x),T_2(y))+\sum r(k)\mid k\in(V_1(x)-V_1(x))+\sum r(\tilde{k})\mid \tilde{k}\in(E(x)-\tilde{E}(x))\}\end{cases}$$

・・（式７）

【００７１】

同様に、上記式２について、頂点ｘが葉（Ｃｈ（ｘ）＝ＮＵＬＬ：空集合）である場合には、明らかに、上記式２の右側の第２項を計算する必要がないため、式８を用いて距離を求めることができる。

また、上記式２について、頂点ｙが葉（Ｃｈ（ｙ）＝ＮＵＬＬ：空集合）である場合には、明らかに、上記式２の右側の第３項を計算する必要がないため、式９を用いて距離を求めることができる。

【００７２】

【数１１】

$$D(\tilde{T}_1(x),\tilde{T}_2(y))=\min\begin{cases}\delta(x,y)+\delta(\tilde{x},\tilde{y})+D(\tilde{F}_1(x),\tilde{F}_2(y)),\\ \min_{y\in Ch(y)}\begin{cases}\min\{D(T_1(x),T_2(y))+\delta(\tilde{x},\tilde{y}),\ D(\tilde{T}_1(x),\tilde{T}_2(y))\}\\ +\sum q(k)\mid k\in(\tilde{V}_2(y)-\tilde{V}_2(y))+\sum q(\tilde{k})\mid \tilde{k}\in(\tilde{E}_2(y)-\tilde{E}_2(y))\end{cases}\end{cases}$$

・・（式８）

【００７３】

【数１２】

$$D(\tilde{T}_a(x),\tilde{T}_b(y))=\min\begin{cases}\delta(x,y)+\delta(\tilde{x},\tilde{y})+D(\tilde{F}_a(x),\tilde{F}_b(y)),\\ \min_{b\in Ch(x)}\begin{cases}\min\{D(T_a(x),T_b(y))+\delta(\tilde{x},\tilde{y}),\quad D(\tilde{T}_a(x),\tilde{T}_b(y))\}\\ +\sum r(k)\mid k\in(\tilde{V}_a(x)-\tilde{V}_a(x))+\sum r(\tilde{k})\mid \tilde{k}\in(\tilde{E}_a(x)-\tilde{E}_a(x))\end{cases}\end{cases}$$

・・（式９）

【００７４】

同様に、上記式３について、頂点ｘが葉（Ｃｈ（ｘ）＝ＮＵＬＬ：空集合）である場合には、明らかに、上記式３の右側の第２項を計算する必要がないため、式１０を用いて距離を求めることができる。

また、上記式３について、頂点ｙが葉（Ｃｈ（ｙ）＝ＮＵＬＬ：空集合）である場合には、明らかに、上記式３の右側の第３項を計算する必要がないため、式１１を用いて距離を求めることができる。

【００７５】

【数１３】

$$D(\tilde{T}_a(x),T_b(y))=\min\begin{cases}\delta(x,y)+\tilde{r}+D(\tilde{F}_a(x),\tilde{F}_b(y)),\\ \min_{y\in Ch(y)}\{D(\tilde{T}_a(x),\tilde{T}_b(y))+\sum q(k)\mid k\in(V_b(y)-V_b(y))+\sum q(\tilde{k})\mid \tilde{k}\in(E_b(y)-\tilde{E}_b(y))\}\end{cases}$$

・・（式１０）

【００７６】

【数１４】

$$D(\tilde{T}_a(x),T_b(y))=\min\begin{cases}\delta(x,y)+\tilde{r}+D(\tilde{F}_a(x),\tilde{F}_b(y)),\\ \min_{b\in Ch(x)}\{D(\tilde{T}_a(x),T_b(y))+\sum r(k)\mid k\in(V_a(x)-V_a(x))+\sum r(\tilde{k})\mid \tilde{k}\in(\tilde{E}_a(x)-\tilde{E}_a(x))\},\end{cases}$$

・・（式１１）

【００７７】

同様に、上記式４について、頂点ｘが葉（Ｃｈ（ｘ）＝ＮＵＬＬ：空集合）である場合には、明らかに、上記式４の右側の第２項を計算する必要がないため、式１２を用いて距離を求めることができる。

また、上記式４について、頂点ｙが葉（Ｃｈ（ｙ）＝ＮＵＬＬ：空集合）である場合には、明らかに、上記式４の右側の第３項を計算する必要がないため、式１３を用いて距離を求めることができる。

【００７８】

【数１５】

$$D(T_i(x),\tilde{T}_k(y))=\min\begin{cases}\delta(x,y)+\tilde{q}+D(\tilde{F}_i(x),\tilde{F}_k(y)),\\ \min_{y\in Ch(y)}\{D(T_i(x),\tilde{T}_k(y))+\sum q(k)\mid k\in(V_k(y)-V_k(y))+\sum q(\tilde{k})\mid \tilde{k}\in(E_k(y)-\tilde{E}_k(y))\}\end{cases}$$

・・（式１２）

【００７９】

【数１６】

$$D(T_i(x),\tilde{T}_k(y))=\min\begin{cases}\delta(x,y)+\tilde{q}+D(\tilde{F}_i(x),\tilde{F}_k(y)),\\ \min_{y\in Ch(x)}\{D(\tilde{T}_i(x),\tilde{T}_k(y))+\sum r(k)\mid k\in(V_i(x)-V_i(x))+\sum r(\tilde{k})\mid \tilde{k}\in(E_i(x)-\tilde{E}_i(x))\}\end{cases}$$

・・（式１３）

【００８０】

次に、上記した特願２００２－０７１２７４号に記載された頂点と辺の対応関係に基づくＲ木間の距離を求める方法を示す。

まず、Ｒ木の根から深さ優先順で頂点と辺に番号を付ける。大きい番号を根とするＲ木の部分から小さい番号を根とするＲ木の部分の順で部分木間の距離を求めて、最後に全体的なＲ木間の距離を求める。

【００８１】

【数１７】

本方法では、上記図６（ａ）に示されるような例えばＲ木である２つの部分木$T_a(x)$、$T_b(y)$の間の距離$D(T_a(x),T_b(y))$を上記式１で求めることができ、上記図６（ｂ）に示されるような例えばＲ木である２つの部分木$\tilde{T}_a(x)$、$\tilde{T}_b(y)$の間の距離$D(\tilde{T}_a(x),\tilde{T}_b(y))$を上記式２で求めることができ、上記図６（ｃ）に示されるような例えばＲ木である２つの部分木$\tilde{T}_a(x)$、$T_b(y)$の間の距離$D(\tilde{T}_a(x),T_b(y))$を上記式３で求めることができ、上記図６（ｄ）に示されるような例えばＲ木である２つの部分木$T_a(x)$、$\tilde{T}_b(y)$の間の距離$D(T_a(x),\tilde{T}_b(y))$を上記式４で求めることができる。

ここで、順序がない森間の距離$D(\tilde{F}_a(x),\tilde{F}_b(y))$、全ての部分木間の距離$D(T_a(x_i),T_b(y))$、$D(\tilde{T}_a(x_i),\tilde{T}_b(y))$、$D(\tilde{T}_a(x_i),T_b(y))$、$D(T_a(x_i),\tilde{T}_b(y))$、$D(T_a(x),T_b(y_j))$、$D(\tilde{T}_a(x),\tilde{T}_b(y_j))$、$D(\tilde{T}_a(x),T_b(y_j))$、$D(T_a(x),\tilde{T}_b(y_j))$は既に求められたものとする。

【００８２】

式１４を用いて、上記図７に示されるような例えば順序がない２つの森の間の距離を求めることができる。

【００８３】

【数１８】

同図に示されるような２つの順序がない森$\tilde{F}_a(x)$、$\tilde{F}_b(x)$の間の距離$D(\tilde{F}_a(x),\tilde{F}_b(y))$が次のように求められる。

$$D(\tilde{F}_a(x),\tilde{F}_b(y))=\sum_{x\in Ch(x)}\left(\sum r(k)\mid k\in V_a(x)+\sum r(\tilde{k})\mid \tilde{k}\in \tilde{E}_a(x)\right)+\sum_{y\in Ch(y)}\left(\sum q(k)\mid k\in V_b(y_j)+\sum q(\tilde{k})\mid \tilde{k}\in \tilde{E}_b(y_j)\right)-W(M_{max})$$

・・（式１４）

【００８４】

ここで、上記式１４中のＷ（M_{max}）は、図８に示すような２部グラフＧ（Ｘ，Ｙ，Ｅ）の最大マッチングの重みである。

また、２部グラフＧ（Ｘ，Ｙ，Ｅ）の頂点x_i（∈Ｘ）と頂点y_j（∈Ｙ）の間の辺ｅ（x_i，y_j）の重みｗ（ｅ（x_i，y_j））を式１５のように設定する。２部グラフＧ（Ｘ，Ｙ，Ｅ）の最大マッチングの重みは、辺ｅ（x_i，y_j）の重みｗ（ｅ（x_i，y_j））の和の最大値に相当する。

【００８５】

【数１９】

$$w(e(x_i,y_j))=\sum r(k)\mid k\in V_a(x_i)+\sum r(\tilde{k})\mid \tilde{k}\in \tilde{E}_a(x_i)+\sum q(k)\mid k\in V_b(y_j)+\sum q(\tilde{k})\mid \tilde{k}\in \tilde{E}_b(y_j)-D(\tilde{T}_a(x_i),\tilde{T}_b(y_j))$$

・・（式１５）

なお、２部グラフＧ（Ｘ，Ｙ，Ｅ）の頂点x_i（∈Ｘ）は順序がない森$\tilde{F}_a(x)$を構成する部分木$\tilde{T}_a(x_i)$（x_i∈Ｃｈ（ｘ））を表し、２部グラフＧ（Ｘ，Ｙ，Ｅ）の頂点y_j（∈Ｙ）は順序がない森$\tilde{F}_b(y)$を構成する部分木$\tilde{T}_b(y_j)$（y_j∈Ｃｈ（ｙ））を表す。

【００８６】

以上のような方法により、ＲＯ木間或いはＲ木間の距離Ｄ（T_a，T_b）＝Ｄ

（T_a（x＝1），T_b（y＝1））を求めることができる。

次に、意味内容比較部８は、式１６或いは式１７を用いて、テキスト文間の距離を求める。

ここで、D（S_1，S_2）は文S_1と文S_2との間の距離を表し、木T_1は文S_1の木構造（ＲＯ木或いはＲ木）を表し、木T_2は文S_2の木構造（ＲＯ木或いはＲ木）を表し、D（T_1，T_2）は木T_1と木T_2との間の距離を表す。

【００８７】

【数２０】

$$D(S_1, S_2) = D(T_1, T_2) \quad \cdots (式16)$$

【００８８】

【数２１】

$$D(S_1, S_2) = \frac{D(T_1, T_2)}{|T_1| + |T_2|} \quad \cdots (式17)$$

【００８９】

次に、具体的な例を用いて、本発明の一実施例に係るテキスト文の意味内容の比較装置及び比較方法の動作を説明する。

本発明の一実施例に係るテキスト文の意味内容の比較装置を用いて、文Ａ「妻の花子は風邪を引きました」と文Ｂ「妻は風邪を引きました」の間の類似度（或いは、相違度）を求める過程と結果を示す。本例では、単語と格の脱落重み及び挿入重みを７０と設定し、単語間の置換重みを１００と設定し、格間の置換重みを１００と設定する。

【００９０】

まず、文Ａと文Ｂを形態素解析した後に、構文意味解析を行い、これにより、これら２つの文Ａ、Ｂがそれぞれ図９（ａ）、図９（ｂ）に示すような例えば根があり順序がある木T_A、T_Bに変換される。

次に、上記式１を用いて変換された２つのＲＯ木間の距離を計算し、最後に、

上記式１６或いは上記式１７を用いて２つのテキスト文Ａ、テキスト文Ｂの間の距離を求める。

【００９１】

上記式１６を用いた場合には、テキスト文Ａ、Ｂ間の距離はＤ（Ａ，Ｂ）＝１４０となり、上記式１７を用いた場合には、テキスト文Ａ、Ｂ間の距離はＤ（Ａ，Ｂ）＝２０（＝１４０／７）となる。ここで、２つのＲＯ木T_A、T_B間の距離はＤ（T_A，T_B）＝１４０であり、２つのＲＯ木T_A、T_Bの頂点の総数は７である。

【００９２】

図１０には、距離Ｄ（T_A，T_B）を与えるＲＯ木間の写像の一つを示している。同図に示されるように、２つのＲＯ木T_A、T_B間の距離は、単語「花子」を脱落するのに必要な脱落重みである７０と、格「ＡＤＪＵＮＣＴ」を脱落するのに必要な脱落重みである７０との和となっている。

【００９３】

以上のように、本発明に係るテキスト文比較装置及び比較方法では、テキスト文を形態素解析し、意味解析を行い、解析されたテキスト文の全体の構文構造と意味をグラフ理論上のＲＯ木或いはＲ木で表現し、つまり、テキスト文の全体の構文構造と意味をＲＯ木或いはＲ木に変換し、テキスト文に出現した単語情報（単語の属性を含む）と単語間の係り受け関係（格）情報をそれぞれＲＯ木或いはＲ木の頂点と辺に格納し、頂点と辺の対応関係に基づくＲＯ木間或いはＲ木間の距離をテキスト文間の意味内容の相違を測る距離に適用して、ＲＯ木間或いはＲ木間の距離を用いてテキスト文間の意味内容の相違を比較することにより、入力された２つのテキスト文間の意味内容を高精度に且つ実時間で求めることができる。

【００９４】

具体的には、本発明では、テキスト文間の距離を、テキスト文間の単語情報の違い、格情報の違い、及び文全体の構造上の違いによって定義したため、本発明に係る距離関数は次の３つの良い性質を有している。（１）意味が似ている且つ構造が似ている２つの文間の距離が非常に小さく評価される。（２）意味が異な

る且つ構造が似ていない２つの文間の距離が非常に大きく評価される。（３）意味が異なるが、構造が似ている２つの文間の距離が単語情報の違いと格情報の違いによって評価される。これにより、２つのテキスト文間の距離を高精度に求めることができる。

【００９５】

また、本例では、ＲＯ木については、木の頂点の数ｎの２乗のオーダー（Ｏ（n^2））で計算することが可能であり、また、Ｒ木については、木の頂点の数ｎの２乗と木の最大の子供の数ｍのオーダー（Ｏ（mn^2））で計算することが可能である。このように実時間での計算が可能である。

【００９６】

ここで、本発明の構成としては、必ずしも以上に示したものに限られず、種々な構成が用いられてもよい。なお、本発明は、例えば本発明に係る方法を実現するためのプログラムなどとして提供することも可能である。

また、本発明の適用分野としては、必ずしも以上に示したものに限られず、本発明は、種々な分野に適用することが可能なものである。

【００９７】

また、本発明において行われる各種の処理としては、例えばプロセッサやメモリ等を備えたハードウエア資源においてプロセッサがＲＯＭ（Read Only Memory）に格納された制御プログラムを実行することにより制御される構成が用いられてもよく、また、例えば当該処理を実行するための各機能手段が独立したハードウエア回路として構成されてもよい。

また、本発明は上記の制御プログラムを格納したフロッピー（登録商標）ディスクやＣＤ（Compact Disc）－ＲＯＭ等のコンピュータにより読み取り可能な記録媒体や当該プログラム（自体）として把握することもでき、当該制御プログラムを記録媒体からコンピュータに入力してプロセッサに実行させることにより、本発明に係る処理を遂行させることができる。

【００９８】

【発明の効果】

以上説明したように、本発明に係るテキスト文比較装置やテキスト文比較方法

によると、テキスト文の全体の構文と意味をグラフ理論上のＲＯ木或いはＲ木で表現し、頂点と辺の対応関係に基づくＲＯ木間の距離或いは頂点と辺の対応関係に基づくＲ木間の距離を用いてテキスト文間の意味内容の相違を比較することにより、入力された２つのテキスト文間の意味内容を高精度に且つ実時間で求めることができる。本発明により、例えば、ドキュメントの意味内容の比較や、意味内容によるドキュメントの分類だけではなく、ユーザの情報探索意図の理解なども可能になる。つまり、自然言語で表現したユーザの要求を、事前に学習して構築されたデータベースの中の内容と比較して、ユーザの情報探索意図を推定することができる。

【図面の簡単な説明】

【図１】　本発明の一実施例に係るテキスト文間の意味内容の比較装置の構成例を示す図である。

【図２】　本発明に係るテキスト文間の意味内容の比較装置及び比較方法を情報端末装置に適用した場合の構成例を示す図である。

【図３】　形態素解析部による解析結果の一例を示す図である。

【図４】　木構造の表現の一例を示す図である。

【図５】　格カテゴリ間の距離の表（リスト）のデータ構造の一例を示す図である。

【図６】　ＲＯ木或いはＲ木からなる２つの部分木の一例を示す図である。

【図７】　ＲＯ木或いはＲ木からなる２つの森の一例を示す図である。

【図８】　２部グラフの一例を示す図である。

【図９】　文Ａと文Ｂの木構造を示す図である。

【図１０】　文Ａと文ＢのＲＯ木間の距離を与える写像の一例を示す図である。

【符号の説明】

１、２１・・外部記憶装置、　２・・形態素解析部、　３・・意味解析部、

４・・木構造変換部、　５・・単語写像重み計算部、

６・・格写像重み計算部、　７・・距離計算部、　８・・意味内容比較部、

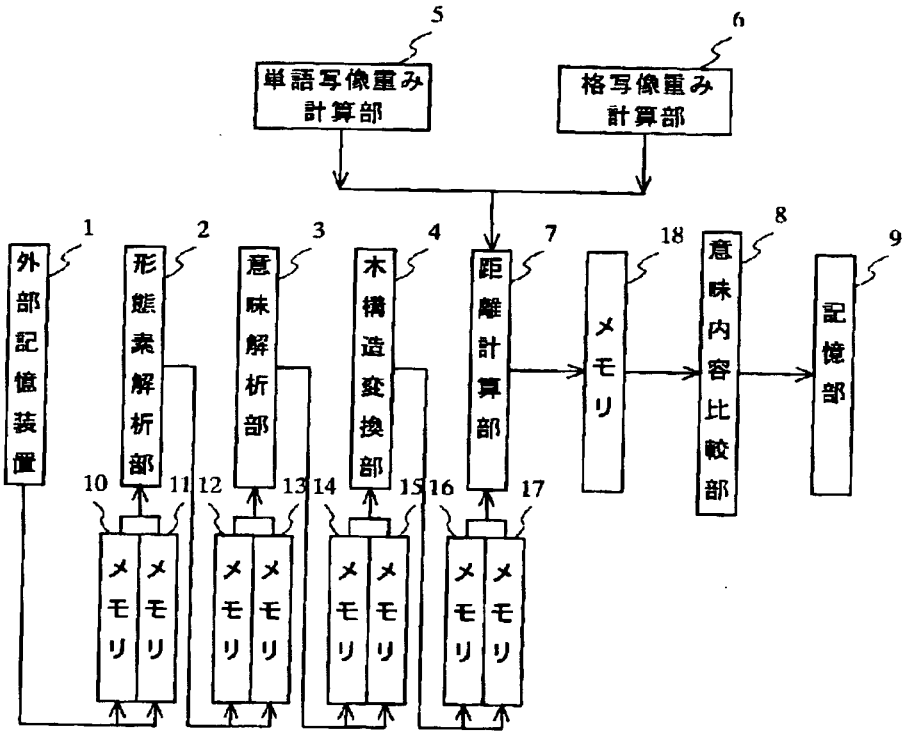
９・・記憶部、　１０～１８・・メモリ、　２０・・情報端末装置、

２２・・キーボード、　２３・・ディスプレイ、　２４・・プロセッサ部、

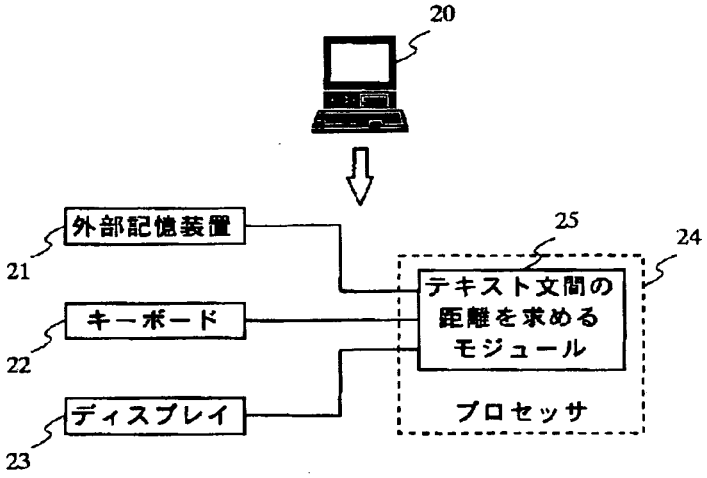
２５・・モジュール、

【書類名】　　図面

【図１】


5
単語写像重み計算部
6
格写像重み計算部
1
外部記憶装置
2
形態素解析部
3
意味解析部
4
木構造変換部
7
距離計算部
18
メモリ
8
意味内容比較部
9
記憶部
10
11
12
13
14
15
16
17
メモリ

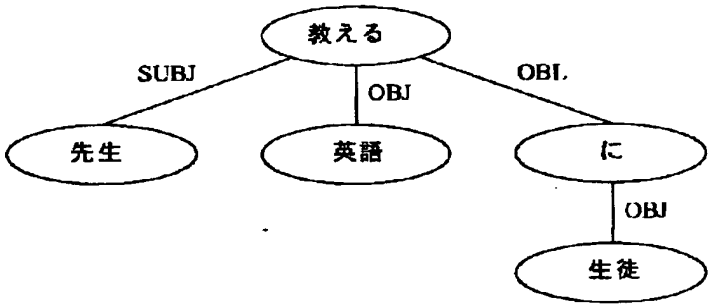
【図２】


20
外部記憶装置
21
キーボード
22
ディスプレイ
23
25
24
テキスト文間の距離を求めるモジュール
プロセッサ

【図３】

先生	先生	センセイ	センセイ	名詞－一般
は	は	ハ	ワ	助詞－係助詞
生徒	生徒	セイト	セイト	名詞－一般
に	に	ニ	ニ	助詞－格助詞－一般
英語	英語	エイゴ	エイゴ	名詞－一般
を	を	ヲ	ヲ	助詞－格助詞－一般
教える	教える	オシエル	オシエル	動詞－自立　一段　基本形
EOS				

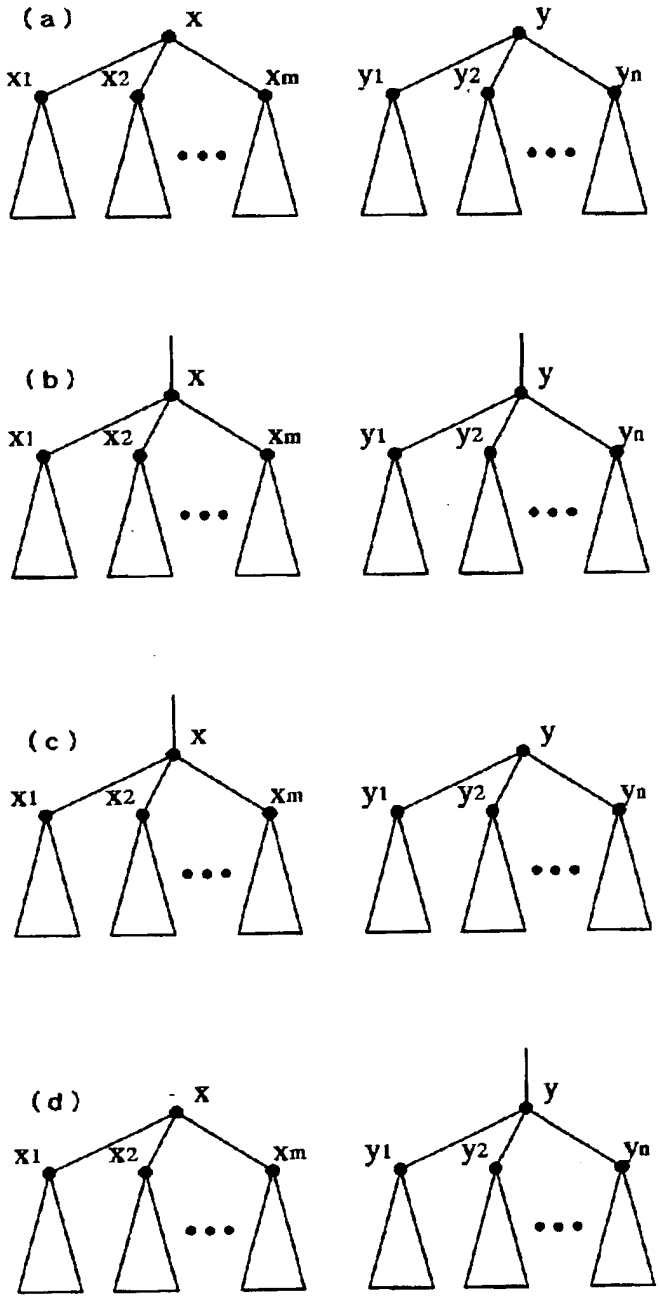
【図４】



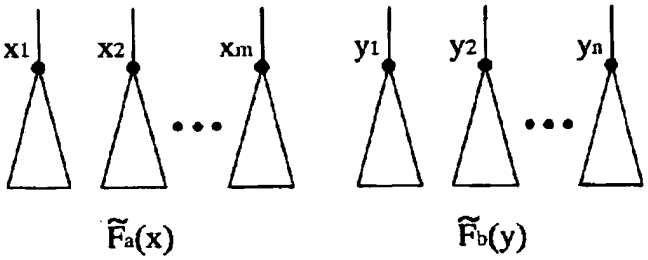
【図５】

格カテゴリ１	格カテゴリ１	距離値11
格カテゴリ１	格カテゴリ２	距離値12
・・・	・・・	
格カテゴリ１	格カテゴリｍ	距離値1m
・・・	・・・	
格カテゴリｍ	格カテゴリ１	距離値m1
格カテゴリｍ	格カテゴリ２	距離値m2
・・・	・・・	
格カテゴリｍ	格カテゴリｍ	距離値mm

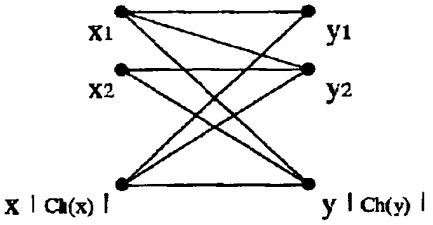
【図６】


(a)
x
x1
x2
xm
y
y1
y2
yn
(b)
x
x1
x2
xm
y
y1
y2
yn
(c)
x
x1
x2
xm
y
y1
y2
yn
(d)
x
x1
x2
xm
y
y1
y2
yn

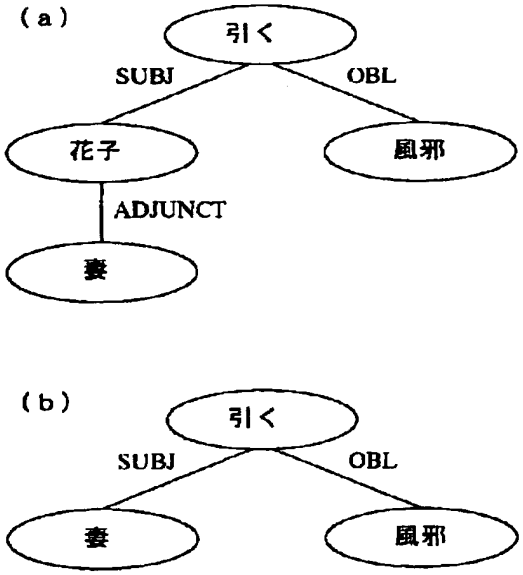
【図７】


x1
x2
xm
y1
y2
yn
F̃a(x)
F̃b(y)

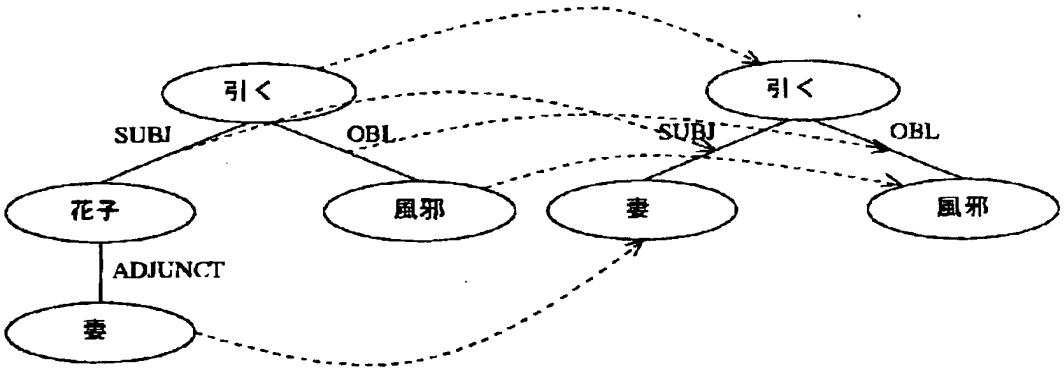
【図８】


x1
x2
y1
y2
x | Ch(x) |
y | Ch(y) |

【図９】


(a)
引く
SUBJ
OBL
花子
風邪
ADJUNCT
妻
(b)
引く
SUBJ
OBL
妻
風邪

【図１０】



【書類名】　　要約書

【要約】

【課題】　高精度に実時間でテキスト文間の比較を行うテキスト文比較装置を提供する。

【解決手段】　木表現手段４が比較対象となるテキスト文をグラフ理論上の根がある木で表現し、情報付与手段４が木表現手段４により表現される木の各頂点に単語情報を付与して各辺に単語間の係り受け関係情報である格情報を付与し、木間距離定義手段７が頂点間の対応関係及び辺間の対応関係に基づく木の間の距離を定義し、木間距離取得手段７が比較対象となるテキスト文の木間について木間距離定義手段７により定義される木間の距離を求め、木間距離適用手段８が木間の距離をテキスト文間の相違を表す距離に適用し、木間距離取得手段８が木間距離適用手段８による適用に基づいて比較対象となるテキスト文間の距離を求める。

【選択図】　　図１

出証特２００３－３０３６２４５

出 願 人 履 歴 情 報

識別番号　〔０００００５４９６〕

１．変更年月日　１９９６年　５月２９日
［変更理由］　住所変更
住　所　東京都港区赤坂二丁目１７番２２号
氏　名　富士ゼロックス株式会社